GOOD TO BE BAD? DISTINGUISHING BETWEEN POSITIVE AND NEGATIVE CITATIONS IN SCIENTIFIC IMPACT

Diana C. Cavalcanti and Ricardo B. C. Prudêncio
Centro de Informática
Universidade Federal de Pernambuco, UFPE
Recife, Brazil
Email: {dcc2, rbcp}@cin.ufpe.br

Shreyasee S. Pradhan, Jatin Y. Shah and Ricardo S. Pietrobon
Member, Research on Research Group
Department of Surgery, Duke University
Durham, USA
Email: {sp133, jys4, rpietro}@duke.edu

Abstract—The impact of a publication is often measured by the number of citations it received, this number being taken as a proxy for the relevance of published work. However, a higher citation index does not necessarily mean that a publication necessarily had a positive feedback from citing authors, as a citation can represent a negative criticism. In order to overcome this limitation, we used sentiment analysis to rate citations as positive, neutral or negative. Adjectives are initially extracted from the citations, with the SentiWordNet lexicon being used to rate the degree of positivity and negativity for each adjective. Relevance scores were then computed to rank citations according to the sentiment expressed in the text corresponding to each citation. As expected for accurate information retrieval systems, higher precision rates were observed in the initial points of the curve. The SRC (0.6728) computed using number of raw citations is lower than the SRC (0.7397) observed by the ranking generated using sentiment scores (Table 3). Conclusion: This result indicates that child articles with higher values of relevance score were in general the ones expressing positive opinion about their parents. Therefore, the ranking generated by sentiment scores had an improved accuracy.

Keywords: Impact Factor, Sentiment Analysis, SentiWordNet, Spearman Ranking Correlation.

I. INTRODUCTION

Journal impact factor is an integral part of the decision making process for most institutional as well as governmental research policies [1]. The citation index reflects the frequency with which the journal's articles are cited in the scientific literature. Conceptually developed in the 1960s, the impact factor has gained acceptance as a quantitative measure of journal quality, having been used to evaluate the quality of journals and associated publications for decades. Authors and academic institutions are now frequently judged and funded simply on the basis of publications in a high impact journal [1]. It is also used by librarians in selecting journals for their collections.

Despite being widely used as a quality indicator for individual researchers and institutions, the impact factor is frequently criticized as having a high degree of uncertainty [2]. For example, a higher citation is no assurance of a positive response from peers in relation to a given published article [3], any two papers possibly having the same number of citations, but with different completely opposite considerations in relation to their quality and value. Despite this criticism, to our knowledge no previous publication has addressed this issue by automatically differentiating between citations with a positive versus a negative tone in relation to the cited article.

In our current work, we investigated the use of sentiment analysis to identify subjectivity in scientific citations. Sentiment Analysis explores the computational study of opinions, feelings and emotions expressed in unstructured sources as texts [4]. The most common goal is to discover what is expressed in the document and classify it into negative, positive or neutral sentiments. Sentiment Analysis has been widely used to process opinions about services and enterprises in social media monitoring systems [5]. Sentiment analysis techniques have also been applied in a variety of domains including the analysis of political discourses, classification of movies, ranking of products among others [6].

The objective of this article is therefore to describe a new application of sentiment analysis in refining the citation impact of a scientific publication, as well as the results obtained on a corpus of experiments. The remaining of this paper is organized as follows. Section II presents the methods developed in this work, followed Section III which describes the corpus of biomedical articles. Section IV provides a brief introduction to sentiment analysis. Section V presents how Sentiment Analysis was applied in our work to compute relevance scores for scientific citations, followed by Section VI which presents the results of our experiments. Finally, Section V concludes the paper with a discussion of our results along with recommendations for future work.

II. METHODS

We built a classifier using the SentiWordNet lexical resource for classifying sentiments on citations from
scientific articles. Construction of the classifier was divided into three steps: (1) selection and extraction belonging to grammatical adjective classes (2) association of negative or positive score for each term calculated from information of positivity and negativity given by SentiWordNet, and (3) classification of the polarity of the document through the average of positive and negative terms for each document. We then ranked each citation with respect to the final score obtained. The citation with highest score was ranked one and accordingly other citations were ranked in the ascending order of the final score. We considered the citation with the highest score to have a high impact while the one with lowest score to have a low impact.

In the following sections, we describe a use case explaining how this system can be integrated into applications evaluating research impact, our data retrieval strategy for the selected sample, how we created a validation set serving as a gold standard for the testing of our algorithms, the algorithm description, and its corresponding performance evaluation.

A. General use case

A set of articles corresponding to a researcher, a team of researchers, institutions, or a geopolitical area such as a country is fed to the system. A citation analysis of each article (referred to as “parent” article) is conducted against a citation database such as Google Scholar (available through an application programming interface - API) to identify the citing article (referred to as “child” article). Each child article is then automatically compared against a repository of full-text articles under open access or other licenses. The system then determines which of the children are available within those repositories. For the child articles found within the repositories, their full-text is searched for the location where the parent is cited and the sentiment in relation to the parent is classified as positive, negative or neutral (neither positive nor negative). For child articles not found in full-text repositories, the sentiment is classified as unknown.

Our study validates the sentiment algorithm to classify each paper into negative, positive, or neutral.

III. DATA RETRIEVAL AND SELECTION OF GOLD STANDARD SUBSET

We randomly selected articles from British Medical Journal (BMJ) from the years 2003 and 2004. BMJ was selected since it is a high impact biomedical journal referred by large number of clinical practitioners; also having a large number of articles publicly available in full-text format (http://www.bmj.com/content/by/year). To create the gold standard subset against which the algorithms were subsequently tested, we initially populated a spreadsheet with the citations of all research articles from BMJ published during the years 2003 and 2004. We then generated a random sample using R statistical language (http://www.r-project.org/).

Our random selection resulted in a total of 37 articles from 2003 and 31 articles from 2004. From the total number of articles (n = 48) we only analyzed only 31 articles. We then searched for all articles citing these 31 articles using Google Scholar (scholar.google.com/, last accessed August 2011). In the remaining of our articles (n=17), either the child papers did not refer the parent at all even if the child was identified through Google Scholar or at times we were unable to retrieve the child articles of the parent. All parent and child articles were then populated in a spreadsheet for further analysis.

A. Validation set

Two researchers (SP and JS) independently tagged all child articles into three different sentiment states, namely positive, negative and neutral. A child article was tagged as positive when it discussed the parent article using terms such as “remarkable,” “first adequately powered”. For example: “Despite some adverse effects, this study was the first adequately powered randomized controlled trial that supported the use of an opioid for the symptomatic relief of dyspnea.” [9] The child article was tagged as negative when it criticized the parent article or highlighted its limitations. For example: “Their findings were conflicting, showing either no reduction in incidence of colorectal or all cancers.” [10] The child article was tagged as neutral when no sentiment was apparent. For example: “Two other trials have also published data on supplementation of these nutrients and the incidence of all cancers, but in both studies the assessment of cancer outcomes was a secondary objective.” [10] (Examples also provided in Table I)

Disagreements among raters were resolved by discussion and mutual consensus.

Below we first define some central concepts related to Sentiment Analysis, followed by an example that illustrates each concept.

IV. SENTIMENT ANALYSIS

Opinion mining or sentiment analysis explores the computational study of opinions, feelings and emotions expressed in unstructured sources such as free text [4]. Its goal is not to determine the topic or theme evaluated in a document such as what is performed in conventional techniques of text classification, but is instead focused on discovering the intended opinion expressed in the document while classifying its polarity [11], [5].

In the business world, where sentiment analysis has been mostly used, the perspective from customers coming in the form of positive and negative comments about products can assist marketing personnel in determining strategic planning. In scientific articles, an article can cite another article in a positive, negative, or neutral form. These statements can be directed at its degree of acceptance regarding experiments, data, statements, methods, or results discussed in the document. As such, sentiment analysis can be framed as a method to evaluate the perceived quality of the publication.
A. SentiWordNet

We used the SentiWordNet [16] version 3.0 lexical resource, developed by Esuli and Sebastiani (2006) [12], which has 117,374 annotated synsets from WordNet 2.0 with sentiments scores. A synset in the original WordNet contains a set of synonyms representing a concept, a grammatical class, and a definition gloss. In SentiWordNet, each synset was augmented with three numerical scores, Pos(s), Neg(s) and Obj(s). These indicated the degree to which the synset was positive, negative or objective in relation to its content. Each of the scores ranged from 0.0 to 1.0 with the overall sum of all three scores adding to 1, specifically: Obj(s) + Pos(s) + Neg(s) = 1

B. Example

Figure 1 illustrates an example synset from SentiWordNet. In this example, the synset \( s \) = \{WICKED, TERRIBLE, SEVERE\}, has an associated definition gloss "intensely or extremely bad or unpleasant in degree or quality", and is associated with the adjective class. The three sentiment scores associated to this synset are Pos(s) = 0, Neg(s) = 0.875 and Obj(s) = 0.125, indicating that this specific synset expresses a predominantly negative sentiment (Figure 1).

![Example synset stored in SentiWordNet](http://sentiwordnet.isti.cnr.it/)

**Figure 1. Example of synset stored in SentiWordNet (figure adapted from the SentiWordNet interface on http://sentiwordnet.isti.cnr.it/)**

V. COMPUTING RELEVANCE SCORES USING SENTIWORDNET

In our sample, each input text i.e., a citation, was associated with a relevance score indicating how positive or negative the citation was in relation to the article being cited. We computed the relevance score by using the following three steps:

1) Extract the adjectives, comparative adjectives and superlative adjectives from the input text by using a Stanford POS-tagger [17].

2) Retrieve all synsets associated with the extracted adjectives, then save the positive score of each retrieved synset.

3) Aggregate the positive scores in order to compute the final relevance score associated with the input text.

While working on step (1), we primarily focused on extracting adjectives from the input since adjectives are usually considered by many authors as good indicators of sentiment on texts [13]. In the statistics provided in Jindal N and Liu B [14], adjectives represent the part of speech with the highest level of non-objective scores. In step (2) we then collected scores for each extracted adjective as we aimed at ranking the child articles according to their sentiment in relation to their parents. Of importance, a single adjective may be associated with more than one synset in the SentiWordNet, i.e., an adjective may have different meanings. Instead of identifying the correct meaning of each adjective, we retrieved the scores of all senses.

In step (3) we calculated the average using equations (1) and (2). Each synset has a positive and negative score. We subtracted the positive score from negative score for all synsets for each term using the equation (1). Finally, the relevance score of an input text was calculated using equation (2) as the average value of score polarity calculated in the previous step.

\[
\text{Score(\text{term}i)} = \frac{1}{n} \sum_{i=1}^{n} \text{score(\text{term})i}
\]

For example, the word “terrible” is an adjective. SentiWordNet has four senses for this word with their respective positive and negative scores \([0, 0.625], [0, 0.875], [0, 0.875], [0.125, 0.25]\). For each term and corresponding grammar class, subtract the positive score from negative score (Equation 1). As a result we obtain: terrible1 = (0.0 - 0.625) = - 0.625, terrible2 = (0.0 - 0.875) = - 0.875, terrible3 = (0.0 - 0.875) = - 0.875 and terrible4 = (0.125 - 0.25) = - 0.125. The final score is then derived by calculating the average values from equation 1, as displayed in Equation 2.

In equation 2 the score (term) corresponds to the values returned by equation (1) and n is the amount of sense for each adjective. The word “terrible” has four senses with the values [- 0.625, - 0.875, - 0.875, - 0.125]. The final score for the term is \((-0.625 + (-0.875) + (-0.875) + (-0.125))/4\), which is equal to -2.5/4, resulting in a score of -0.625. If the value is below zero, we classify the term as having a negative sentiment. If the value is above zero, then it is classified as having a positive sentiment. And if the value equals zero, then it is classified as neutral. As the score for the word “terrible” is -0.625 which is below zero hence it is classified as negative.

<table>
<thead>
<tr>
<th>Parent Article</th>
<th>Citing Article</th>
<th>Quote about parent</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact of supplementing newborn infants with vitamin A on early infant mortality: community based randomised trial in southern India</td>
<td>Effect of 50 000 IU vitamin A given with BCG vaccine on mortality in infants in Guinea-Bissau: randomised placebo controlled trial</td>
<td>When we started our trial, two similar trials from Asia had reported beneficial effects of vitamin A supplementation at birth.</td>
<td>Good</td>
</tr>
</tbody>
</table>
In the final step, the semantic orientation of each document is calculated using equation 2. Initially a final score was obtained for each term used to express the sentiment by calculating the average of all scores. Then, an overall score was obtained by calculating the average using the final score of all terms that were identified for each document.

Table 1 presents an example of a parent article along with its child article. We used the text from the field “Quote about parent”. From the text of this example, after the adoption of the previous steps, we extracted the adjectives similar and beneficial with their score presented in table II. The resulting score was 0.275 for the word ‘similar’ and 0.625 for the word ‘beneficial’. The scores were then aggregated using equation 2, which returned the value 0.45, ultimately resulting in a positive classification for the citation.

We computed classical metrics for evaluating information retrieval systems, including precision curves, recall and f-measure.

The curve of precision at position n in the ranking is defined as the number of actual positive citations observed up to position n divided by n. Precision in this context measures the ability of ranking in the initial positions just true positive citations.

In figure 2, we present the curve of precision computed from intervals of 10 citations. The precision measure starts with 70% for n=10 and progressively decreases as the number of ranked citations increases. As expected for accurate information retrieval systems, higher precision rates were observed in the initial points of the curve. This result indicates that child articles with higher values of relevance score were in general the ones expressing positive opinion about their parents.

### Table II. Example of Adjective Class with Positive and Negative Score

<table>
<thead>
<tr>
<th>Term</th>
<th>Positive score</th>
<th>Negative score</th>
<th>Score (termsense)</th>
<th>ScoreF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar</td>
<td>0.375</td>
<td>0.25</td>
<td>0.125</td>
<td>(0.125 + 0.25 + (-0.25) + 0.75 + 0.5) / 5 = 0.275</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.25</td>
<td>-0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.875</td>
<td>0.125</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Beneficial</td>
<td>0.625</td>
<td>0</td>
<td>0.625</td>
<td>(0.625 + 0 = 0.625 / 2 = 0.625)</td>
</tr>
<tr>
<td></td>
<td>(0.275 + 0.625)</td>
<td>0</td>
<td>0.9 / 2 = 0.45</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Curve of precision observed for the citations sorted by the relevance scores

Figure 3 presents the curve of recall for our entire sample. Recall at position n is measured in our work as the number of actual positive citations observed up to position n in the ranking divided by the total number of positive citations in the corpus of experiments. Recall measures the amount of relevant items already retrieved in relation to the total amount of relevant items available. As expected, the recall values increase as the number of inspected citations increases. The stable increasing behavior of recall is also a consequence that true positive citations are ordered in general in the top positions considering the relevance score.

### VI. Experiments

In this section, we provide results for our experiments while placing them in context. We evaluated a total of 31 parent papers, corresponding to a total of 140 child citations. Of these, 24 were rated as having a positive sentiment, 8 as having a negative sentiment, and 108 as having a neutral sentiment.

#### A. Ranking citing articles by sentiment score

We initially ranked citations according to their relevance scores and evaluated the quality of the generated ranking.
This result agrees with the common trade-off between precision and recall [15]. In general, in order to achieve a higher level of recall, precision is lowered. In order to analyze both measures, we adopted the F-Measure metric, representing the harmonic mean between precision and recall (see equation 3).

$$F = \frac{2 \times P \times R}{P + R}$$  \hspace{1cm} (3)

Figure 4 presents the f-measure curve observed in our experiments. The highest value for f-measure was observed for n = 50, indicating that higher levels of recall achieved in this position compensated the decrease in precision observed. From that position, the f-measure steadily decreases as the number of citations increases.

### B. Ranking parent articles by the number of positive citations

In the previous section, we evaluated whether the sentiment score could provide useful information to reveal positive citations on relevant papers. Once the positive citations are assigned by following the ranking generated by the sentiment score, one can assign to each parent paper a positive impact factor. This factor can be simply defined as the number of positive citations a parent paper had. In this section, we evaluate the utility of such impact factor to sort the parent papers.

In order to generate a ranking of parent papers the following procedure was adopted:

1) Given the ranking of child citations generated by the sentiment score, we assigned the top 50 ones as positive citations. The number of top 50 citations was chosen based on the highest f-measure observed in the previous experiments (see section 6.1).

2) For each parent paper, we counted the number of citations to that paper which were assigned in the previous step as positive.

3) Finally, we ranked the parent papers by the number of assigned positive citations.

In order to evaluate the quality of the generated ranking, we compared it to the true ranking of parent papers sorted by the true number of positive citations. In this experiment, we measure how correlated is the ranking of parent papers inferred using sentiment scores of citations and the actual ranking of parent papers. For this we adopt the Spearman Ranking Correlation (SRC), defined by the equation:

$$SRC = 1 - \frac{6 \times \sum_{k=1}^{K} (rr_k - ir_k)^2}{K^3 - K}$$  \hspace{1cm} (4)

In this equation (4), $rr_k$ and $ir_k$ are, respectively, the rank of the parent article k in the recommended ranking and the ideal ranking and K is the number of parent articles. SRC assumes values between -1 and 1. Values close to 1 indicate that the two rankings have many agreement positions and values near to -1 indicate disagreement between the rankings.
The interpretation is that the greater the number of citations a paper has, the greater the number of positive citations, which means that it is possible to infer a positive impact factor of a paper just by considering its number of raw citations. However, the relationship between raw citations and actual positive citations is not linear. In fact, by inspecting the data set of experiments we observed that the parent paper with the highest number of positive citations (10 positive citations) was not the most cited paper. Instead, the most cited paper was only the 4th paper concerning positive sentiments. We therefore argue that these two criteria, i.e., number of raw citations and number of positive citations, provide different and synergistic levels of information that can be used to evaluate scientific impact.

Finally, we also computed in our experiments the SRC value by adopting number of raw citations as the criterion for ranking the parent papers. Although high (0.6728), the SRC computed using number of raw citations is lower than the SRC (0.7397) observed by the ranking generated using sentiment scores (Table III). Therefore, the ranking generated by sentiment scores had an improved accuracy.

VII. DISCUSSION

To our knowledge, this is the first article where citation analysis has been expanded to determine the sentiment associated with specific citations. Our analysis reveals that the child articles with higher relevance scores were in general the ones expressing positive opinion about their parents. We also found that the parent article receiving the highest number of positive citations was not the most cited paper. This method can therefore a higher degree of granularity in citation analysis, potentially decreasing the level of criticism associated with this metric. Of importance, we have empirically validated this method to demonstrate that our algorithm can provide reliable estimates of sentiment in relation to the citation metric.

Numerous efforts have attempted to classify articles by their overall sentiments like classifying whether a review is positive or negative [16] [17] [18] [19]. Other researchers have analyzed learning words and phrases having prior positive or negative polarity [20] [24]. These studies differ from ours in a number of ways. Kim and Hovy (2004), Hu and Liu (2004), and Grefenstette et al. (2004) restricted their tags to positive and negative and excluded the neutral category while we included it in our analyses [21] [22] [23].

In the study conducted by Wilson et al. (2009) noted that even if words are positive or negative they are frequently used in neutral context. When these neutral instances are removed the performance of the classification system improves. In contrast to our results, they also highlighted that while classifying the positive and negative polarity, features for negative polarity prove to be the most important [25].

Despite a significant contribution to the literature, our study has limitations. First, to ensure feasibility, we had reduced the citations to articles within BMJ. While we have no reasons to believe that authors publishing in BMJ might express their sentiments in relation to articles that they cite in a way that is different from other general journals, variation might occur depending on country of origin, clinical specialty, and time frame. Finally, although our results demonstrate good reliability, several of our metrics could be potentially improved with alternative algorithms that have not been tested in this project, including for instance the use of supervised machine learning methods.

In conclusion, we have found that sentiment analysis related to citation metrics can be performed in a reliable manner while providing additional details on why a certain paper or author is commonly cited. Future investigations should attempt to apply this methodology on a broader basis. Particularly fruitful experiments could be achieved in the so called green open access repositories, where full-text papers by researchers of a given institution can be compared against citations from large open-access repositories.
REFERENCES